Mirror descent algorithm in stochastic online optimization with noisy first order oracle

Daniil Merkulov

MIPT, Moscow, Russia

bratishka.mipt@gmail.com

Outline

- **Why mirror descent?**
- **Problem formulation**
- **Mirror descent algorithm**
- **Main results**
- **Why is it important**

Problem formulation

Player: chooses $x_k \in Q$ (based on previous experience)

Problem formulation

Player: chooses $x_k \in Q$ (based on previous experience)
Nature: chooses $f_k(x)$(maybe, adversely) from considered class of functions

Problem formulation

Player: chooses $x_k \in Q$ (based on previous experience)
Nature: chooses $f_k(x)$(maybe, adversely) from considered class of functions
The **loss** of player on k-th step:

$$f_k(x_k) - \min_{x \in Q} f_k(x)$$

Problem formulation

Player: chooses $x_k \in Q$ (based on previous experience)
Nature: chooses $f_k(x)$(maybe, adversely) from considered class of functions
The **loss** of player on k-th step:

$$f_k(x_k) - \min_{x \in Q} f_k(x)$$

Average loss in N steps(**regret**):

$$R_N = \frac{1}{N}\sum_{k=1}^{N} f_k(x_k) - \min_{x \in Q} \frac{1}{N}\sum_{k=1}^{N} f_k(x)$$

Problem formulation(Stochastic case)

We need to find a sequence $\{x_k\} \in Q$ that minimizes (Pseudo)Regret:

$$\hat{R}_N = \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}_{\xi_k}\left[f_k(x_k;\xi_k)\right] - \min_{x\in Q}\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}_{\xi_k}\left[f_k(x;\xi_k)\right]$$

Only **noisy subgradient** $g_k(x_k)$ can be obtained from the oracle:

$$\mathbb{E}_{\xi_k}\|\nabla f_k(x_k,\xi_k) - g_k(x_k)\|_* \leq \sigma$$

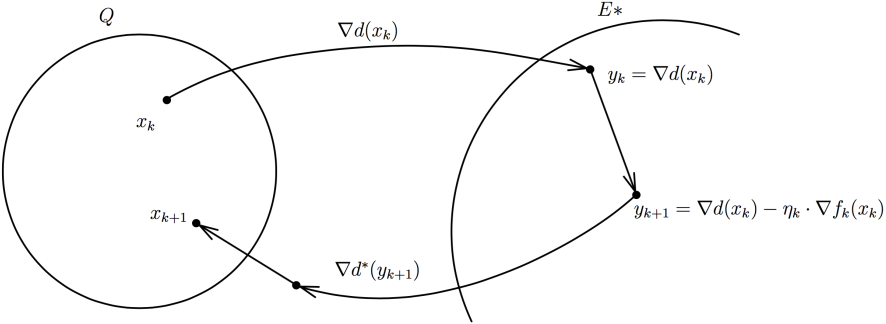
Problem formulation

Full description

$$\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}_{\xi_k}\left[f_k(x_k;\xi_k)\right]-\min_{x\in Q}\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}_{\xi_k}\left[f_k(x;\xi_k)\right]\le\varepsilon$$

- $\mathbb{E}_{\xi_k}\|\nabla f_k(x_k,\xi_k)-g_k(x_k)\|_*\le\sigma$ noisy first order oracle
- $f_1(x,\xi_1),\dots,f_k(x,\xi_k)$ - (strongly) convex functions on x
- $\mathbb{E}_{\xi_k}\|\nabla f_k(x,\xi_k)\|_*\le L$ for all k
- Q - closed convex set in $\mathbb{R}^n$
- $\xi_1,\dots\xi_k$ - i.i.d.

Mirror descent algorithm



$$\begin{cases} y_{k+1} = \nabla d(x_k) - \eta_k \cdot \nabla f_k(x_k) \\ x_{k+1} = \underset{x \in Q}{\mathrm{Proj}}\{\nabla d^{-1}(y_{k+1})\} \end{cases}$$

Mirror descent setup

- norm $\|\cdot\|$ on E ($\|\xi\|_* = \max_x\{\langle \xi, x\rangle : \|x\| \leq 1\}$)
- distance generating function $d(x) : Q \to \mathbb{R}$, which should be:
  - convex and continiously differentable on Q
  - '1'-strongly convex w.r.t. $\|\cdot\|$:

$$\forall x, y \in Q : d(x) \geq d(y) + \langle \nabla d(y), x - y\rangle + \frac{1}{2}\|x - y\|^2$$

- Bregman distance: $\rho(x, y) = d(x) - d(y) - \langle \nabla d(y), x - y\rangle$, by strong convexity of $d(x)$:

$$\rho(x, y) \geq \frac{1}{2}\|x - y\|^2 \quad \forall x, y \in Q$$

- Prox-diameter R of Q:

$$R^2 = \max_{x\in Q} d(x) - \min_{x\in Q} d(x)$$

Lemma

Mirror descent can be rewritten as:

$$x_{k+1} = \underset{x \in Q}{\operatorname{argmin}} \Big\{ \eta_k \langle \nabla f_k(x_k), x \rangle + \beta_k \rho(x, x_k) \Big\}$$

Suggestion

We use following version of MD:

$$x_{k+1} = \underset{x \in Q}{\operatorname{argmin}} \Big\{ \eta_k \langle g_k(x_k), x \rangle + \beta_k \rho(x, x_k) \Big\}$$

Optimality conditions:

$$\eta_k \langle g_k(x_k), x_{k+1} - x \rangle \leq \langle \beta_k \nabla d(x_{k+1}) - \beta_k \nabla d(x_k), x - x_{k+1} \rangle$$

Scheme of proof

$$f_k(x_k) - f_k(x) \overset{①}{\leq} \langle \nabla f_k(x_k), x_k - x \rangle - \frac{\mu}{2}||x_k - x||^2$$

where:
① - (strong)convexity $f_k(x)$

Scheme of proof

$$f_k(x_k) - f_k(x) \overset{①}{\leq} \langle \nabla f_k(x_k), x_k - x\rangle - \frac{\mu}{2}||x_k - x||^2$$
$$\overset{②}{\leq} \langle g_k(x_k), x_k - x\rangle - \frac{\mu}{2}||x_k - x||^2 + \mathbf{2}\sigma R$$

where:
① - (strong)convexity $f_k(x)$,
② - definition of noise

Scheme of proof

$$
\begin{aligned}
f_k(x_k) - f_k(x) &\overset{①}{\leq} \langle \nabla f_k(x_k), x_k - x \rangle - \frac{\mu}{2}||x_k - x||^2 \\
&\overset{②}{\leq} \langle g_k(x_k), x_k - x \rangle - \frac{\mu}{2}||x_k - x||^2 + 2\sigma R \\
&= \langle g_k(x_k), x_k - x_{k+1} \rangle + \langle g_k(x_k), x_{k+1} - x \rangle - \frac{\mu}{2}||x_k - x||^2 + 2\sigma R
\end{aligned}
$$

where:
① - (strong)convexity $f_k(x)$,
② - definition of noise

Scheme of proof

$$
\begin{aligned}
f_k(x_k) - f_k(x) &\overset{①}{\leq} \langle \nabla f_k(x_k), x_k - x \rangle - \frac{\mu}{2}||x_k - x||^2 \\
&\overset{②}{\leq} \langle g_k(x_k), x_k - x \rangle - \frac{\mu}{2}||x_k - x||^2 + 2\sigma R \\
&= \langle g_k(x_k), x_k - x_{k+1} \rangle + \langle g_k(x_k), x_{k+1} - x \rangle - \frac{\mu}{2}||x_k - x||^2 + 2\sigma R \\
&\overset{③}{\leq} \langle g_k(x_k), x_k - x_{k+1} \rangle + \frac{\beta_k}{\eta_k}\langle \nabla d(x_{k+1}) - \nabla d(x_k), x - x_{k+1} \rangle \\
&- \frac{\mu}{2}||x_k - x||^2 + 2\sigma R
\end{aligned}
$$

where:

① - (strong)convexity $f_k(x)$,

② - definition of noise,

③ - optimality conditions

Results(convex case)

Lemma

$\forall x \in Q$ and for $k = 1 \ldots N$ if f_k - convex functions.

$$\mathbb{E}_{\xi_k} f_k(x_k, \xi_k) - \mathbb{E}_{\xi_k} f_k(x, \xi_k) \leq \frac{\beta_k}{\eta_k}\rho(x, x_k) - \frac{\beta_k}{\eta_k}\rho(x, x_{k+1}) + \frac{\eta_k ||g_k(x_k)||_*^2}{2\beta_k} + \sigma \cdot 2R$$

Results(convex case)

Lemma

$\forall x \in Q$ *and for* $k = 1 \dots N$ *if* f_k *- convex functions.*

$$\mathbb{E}_{\xi_k} f_k(x_k, \xi_k) - \mathbb{E}_{\xi_k} f_k(x, \xi_k) \le \frac{\beta_k}{\eta_k}\rho(x, x_k) - \frac{\beta_k}{\eta_k}\rho(x, x_{k+1}) + \frac{\eta_k ||g_k(x_k)||_*^2}{2\beta_k} + \sigma \cdot 2R$$

since $\mathbb{E}_{\xi_k} \|\nabla f_k(x, \xi_k)\|_* \le L$ and $\mathbb{E}_{\xi_k} \|\nabla f_k(x_k, \xi_k) - g_k(x_k)\|_* \le \sigma$

Results(convex case)

Lemma

$\forall x \in Q$ and for $k = 1 \dots N$ if f_k - convex functions.

$$\mathbb{E}_{\xi_k} f_k(x_k, \xi_k) - \mathbb{E}_{\xi_k} f_k(x, \xi_k) \leq \frac{\beta_k}{\eta_k}\rho(x, x_k) - \frac{\beta_k}{\eta_k}\rho(x, x_{k+1}) + \frac{\eta_k ||g_k(x_k)||_*^2}{2\beta_k} + \sigma \cdot 2R$$

since $\mathbb{E}_{\xi_k} \|\nabla f_k(x, \xi_k)\|_* \leq L$ and $\mathbb{E}_{\xi_k} \|\nabla f_k(x_k, \xi_k) - g_k(x_k)\|_* \leq \sigma$

$$\mathbb{E}_{\xi_k} f_k(x_k, \xi_k) - \mathbb{E}_{\xi_k} f_k(x, \xi_k) \leq \frac{\beta_k}{\eta_k}\rho(x, x_k) - \frac{\beta_k}{\eta_k}\rho(x, x_{k+1}) + \frac{\eta_k (L+\sigma)^2}{2\beta_k} + \sigma \cdot 2R$$

Summarize for $k = 1 \dots N$, we get $\hat{R}_N \cdot N$ in the left part:

$$\sum_{k=1}^{N} \mathbb{E}_{\xi_k}\left[f_k(x_k; \xi_k)\right] - \min_{x \in Q} \sum_{k=1}^{N} \mathbb{E}_{\xi_k}\left[f_k(x; \xi_k)\right]$$

Results(convex case)

Choosing stepsize:

$$\eta_k = 1; \qquad \beta_k = \frac{L+\sigma}{R}\sqrt{\frac{N}{2}}$$

Upper bound

$$\hat{R}_N \leq \sqrt{\frac{2R^2(L+\sigma)^2}{N}} + 2\sigma R$$

Results(strongly convex case)

Lemma

$\forall x \in Q$ and for $k = 1 \dots N$ if f_k - strongly convex functions.

$$\mathbb{E}_{\xi_k} f_k(x_k, \xi_k) - \mathbb{E}_{\xi_k} f_k(x, \xi_k) \leq \frac{\beta_k}{\eta_k}\rho(x, x_k) - \frac{\beta_k}{\eta_k}\rho(x, x_{k+1}) + \frac{\eta_k ||g_k(x_k)||_*^2}{2\beta_k} - \frac{\mu}{2}||x_k - x||^2 + \sigma \cdot 2R$$

Results(strongly convex case)

Lemma

$\forall x \in Q$ *and for* $k = 1 \dots N$ *if* f_k *- strongly convex functions.*

$$\mathbb{E}_{\xi_k} f_k(x_k, \xi_k) - \mathbb{E}_{\xi_k} f_k(x, \xi_k) \leq \frac{\beta_k}{\eta_k}\rho(x, x_k) - \frac{\beta_k}{\eta_k}\rho(x, x_{k+1}) + \frac{\eta_k ||g_k(x_k)||_*^2}{2\beta_k} - \frac{\mu}{2}||x_k - x||^2 + \sigma \cdot 2R$$

Choosing stepsize: $\eta_k = \dfrac{1}{\mu k}$, $\beta_k = 1$, $\rho(x, y) = \dfrac{1}{2}||x - y||^2$

Upper bound

$$\hat{R}_N \leq \frac{(L+\sigma)^2}{2\mu} \frac{1 + \ln N}{N} + 2\sigma R$$

Appendix: Unbounded area

Offline convex case

$$\forall x \in Q \quad f(x_k) - f(x) \leq \frac{\beta_k}{\eta_k}\rho(x, x_k) - \frac{\beta_k}{\eta_k}\rho(x, x_{k+1}) + \frac{\eta_k ||g_k(x_k)||_*^2}{2\beta_k}$$

$$x^* = \operatorname{argmin} f(x)_{x\in Q} \quad \sum_{k=0}^{N-1} [\alpha_k(f(x_k) - f(x^*)] \leq \rho(x^*, x_0) - \rho(x^*, x_N) + \Delta_{N-1}$$

$$0 \leq \rho(x^*, x_0) - \rho(x^*, x_N) + \Delta_{N-1}$$

$$\rho(x^*, x_N) \leq \rho(x^*, x_0) + \Delta_{N-1}$$

$$\frac{1}{2}\|x_N - x^*\| \leq \rho(x^*, x_N) \leq \rho(x^*, x_0) + \Delta_{N-1}$$

$$\|x_N - x^*\| \leq 2\rho(x^*, x_0) + 2\Delta_{N-1}$$

Where $\Delta_N = \frac{\sum_{k=0}^{N} \alpha_k^2}{2}\|g_k(x_k)\|_*^2$

Why it is important?

- Computing $\nabla f_k(x)$ may be very expensive

Why it is important?

- Computing $\nabla f_k(x)$ may be very expensive
- Instead of gradient we can use approximation as $g_k(x_k)$!

Why it is important?

- Computing $\nabla f_k(x)$ may be very expensive
- Instead of gradient we can use approximation as $g_k(x_k)$!
- Mirror descent type methods can be applied to a new class of problems (nonsmooth, with zero order noisy oracle).

Thank you for your attention!